Applying Latent Dirichlet Allocation to Group Discovery in Large Graphs

Keith Henderson and Tina Eliassi-Rad
Lawrence Livermore National Laboratory*
Box 808, L-560, Livermore, CA 94551 USA
{keith, eliassi}@llnl.gov

ABSTRACT

This paper introduces *LDA-G*, a scalable Bayesian approach to finding latent group structures in large real-world graph data. Existing Bayesian approaches for group discovery (such as *Infinite Relational Models*) have only been applied to small graphs with a couple of hundred nodes. LDA-G (short for *Latent Dirichlet Allocation for Graphs*) utilizes a well-known topic modeling algorithm to find latent group structure. Specifically, we modify *Latent Dirichlet Allocation* (LDA) to operate on graph data instead of text corpora. Our modifications reflect the differences between real-world graph data and text corpora (e.g., a node's neighbor count vs. a document's word count). In our empirical study, we apply LDA-G to several large graphs (with thousands of nodes) from PubMed (a scientific publication repository). We compare LDA-G's quantitative performance on link prediction with two existing approaches: one Bayesian (namely, *Infinite Relational Model*) and one non-Bayesian (namely, *Cross-associations*). On average, LDA-G outperforms IRM by 15% and Cross-associations by 25% (in terms of area under the ROC curve). Furthermore, we demonstrate that LDA-G can discover useful qualitative information.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Probabilistic algorithms; H.2.8 [**Database Management**]: Database Applications—*Data Mining*; I.2.6 [**Artificial Intelligence**]: Learning; I.5.1 [**Pattern Recognition**]: Models—*Statistical*

General Terms

Algorithms, Design, Performance, Experimentation.

Keywords

Latent Dirichlet allocation, social network analysis, group discovery, graph mining.

1. INTRODUCTION

We address the problem of discovering latent group structure in a large real-world graph, $G = (V, E)$.[1] Despite the recent interest in this problem, computationally scalable algorithms that produce expressive groups (e.g., in terms of probability distribution over group memberships) are rare. Bayesian approaches (such as Infinite Relational Models [5]) produce expressive groups but they have $O(|V|^2)$ space and runtime complexity. Compression-based approaches like Cross-associations [3] are $O(|E|)$,[2] but the groups are not expressive and in some cases not meaningful (see Section 5).

In this paper, we present a scalable Bayesian alternative to the existing approaches, called *Latent Dirichlet Allocation for Graphs* (*LDA-G*). Our approach is an adaptation of the commonly used topic modeling approach called *Latent Dirichlet Allocation* (*LDA*) [2]. We demonstrate that by modifying LDA only slightly, we can apply it to large real-world graph data with excellent quantitative and qualitative results. We demonstrate that it is superior, with respect to link prediction, to a canonical sampling-based Bayesian method (i.e., Infinite Relational Models [5]) and a canonical scalable compression-based method (i.e., Cross-associations [3]). Moreover, we show that it can be used in a practical analysis context to extract useful qualitative information from large graph data.

Our contributions are as follows:

- *We introduce LDA-G, a scalable alternative to existing Bayesian techniques for group discovery in large real-world graphs.*
- *We demonstrate how LDA-G can be applied to real problems of interest in data mining in terms of both quantitative and qualitative results.*

The paper is organized as follows. Section 2 presents a brief overview of existing approaches to which we compare LDA-G, as well as a short explanation of LDA for topic modeling. Section 3 describes the differences between applying LDA to graph data vs. text corpora, and our modifications to LDA. Section 4 outlines our experimental study.

*This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. LLNL-CONF-408036.

[1] In this paper, we use these terms interchangeably: (*i*) graph and network, (*ii*) vertex and node, (*iii*) edge and link.

[2] In real-world graphs, $|E|$ is assumed to be much smaller than $|V|^2$.

SAC'09 March 8-12, 2009, Honolulu, Hawaii, U.S.A.
Copyright 2009 ACM 978-1-60558-166-8/09/03 ...$5.00.

Section 5 reports our quantitative and qualitative results. Lastly, Section 6 provides concluding remarks and promising future directions.

2. RELATED WORK

Bayesian approaches to group discovery:[3] The Infinite Relational Model (IRM) is a canonical form of nonparametric Bayesian approaches to group discovery in graph (i.e. relational) data [5]. It assumes a generative model in which each vertex v_i is assigned a group z_i from an infinite set of groups. The model also contains a matrix η of probabilities, such that the probability of a vertex v_i having a link to a vertex v_j is equal to the entry η_{ij}. The full generative model is as follows:

$$z|\gamma \sim CRP(\gamma) \tag{1}$$

$$\eta(a,b)|\beta \sim Beta(\beta, \beta) \tag{2}$$

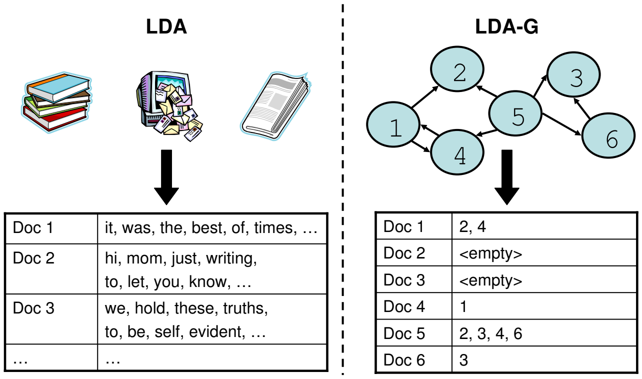
$$R(i,j)|z,\eta \sim Bernoulli(\eta(z_i, z_j)) \tag{3}$$

$R(i,j)$ is a binary value representing the presence or absence of a link. A primary drawback of IRM is its requirement for observations of both present (1-valued) and absent (0-valued) edges in the graph. Thus, in order for IRM to learn a predictive model, it requires space and time proportional to $|V|^2$.

If only present (1-valued) edges are considered, the η matrix entries will all be equal to 1 – i.e., the model will believe that all unobserved edges are present and the graph is a clique. If a sparse representation of the graph is used, then at inference time IRM will believe that any unobserved edges are actually absent, which again biases the model.

On real-world graphs that are sufficiently small (about a couple of hundred nodes), IRM has been demonstrated to provide highly predictive models [5]. Our inference on IRM uses a standard Gibbs sampling approach [12].

Non-Bayesian approaches to group discovery: Generally speaking, non-Bayesian approaches to group discovery in graphs can be divided into two categories: (a) those that rely on compression/MDL such as [3] and (b) those that rely on graph-theory such as [10]. Algorithms based on these two approaches are more scalable when compared with Bayesian approaches. However, their model of a graph's group structure is not as expressive (e.g., in term of probability distribution over group memberships). For instance, Cross-associations [3] uses Minimum Description Length (MDL) to generate a compression-based grouping of vertices. It minimizes the total encoding cost by reordering vertices (source vertices and destination vertices are permuted independently). The total encoding cost is defined by the sum of the code cost and the description cost. The code cost is essentially a measure of the entropy in each discovered block of the reordered adjacency matrix. The description cost measures how many bits it takes to actually describe the groups themselves. Cross-associations' runtime complexity is $O(|E|)$.



Doc 1	it, was, the, best, of, times, …
Doc 2	hi, mom, just, writing, to, let, you, know, …
Doc 3	we, hold, these, truths, to, be, self, evident, …
…	…

Doc 1	2, 4
Doc 2	<empty>
Doc 3	<empty>
Doc 4	1
Doc 5	2, 3, 4, 6
Doc 6	3

Figure 1: Document representation in LDA vs. LDA-G. Significant differences exist between the two document representations such as the length of documents.

Topic modeling: A canonical form of topic modeling is the latent Dirichlet allocation (LDA) [2]. LDA is a hierarchical nonparametric Bayesian approach to topic discovery in text corpora. It assumes a generative model in which each document d_i from a corpus samples (latent) topics from a multinomial distribution $Mult(\theta)$, and each topic z_i samples words in the vocabulary from a multinomial distribution $p(w|z_i)$. Gibbs sampling, such as in [4], can be used to estimate the posterior probability on configurations of the model. A given configuration is an assignment $\mathbf{z} = < z_1, z_2, ..., z_n >$, where each entry corresponds to the topic of a given word in the corpus. The full generative model for a simplified version of LDA is as follows:

$$w_i|z_i, \phi^{(z_i)} \sim Discrete(\phi^{(z_i)}) \tag{4}$$

$$\phi \sim Dirichlet(\beta) \tag{5}$$

$$z_i|\theta^{d_i} \sim Discrete(\theta^{d_i}) \tag{6}$$

$$\theta \sim Dirichlet(\alpha) \tag{7}$$

Our inference on LDA and LDA-G uses a standard Gibbs sampling approach [12]. The runtime complexity of this inference approach is $O(NKM)$. For LDA, N is the number of documents, K is the number of topics, and M is the average length of documents. For LDA-G, N is the number of nodes, K is the number of groups, and M is the average node degree in the graph. The space complexity for both is $O(N(K+M))$.

3. LATENT DIRICHLET ALLOCATION FOR GRAPHS

To use LDA, we represent the input graph as a collection of documents containing words. We assume that the input graph is directed; and if it is undirected, we transform each undirected edge into two directed edges. Each vertex is treated as a document, with the "text" of a document corresponding to the edge-list of the vertex (see Figure 1). In this manner, each vertex also plays a role as a word in the vocabulary. We use the simplifying assumption that the behavior of a vertex as a document is independent of its behavior as a word. The result is a "corpus" of "documents" with "words" that is the input to the LDA model. The characteristics of this corpus is significantly different than a typical text corpus.

There are three primary differences between real-world

[3] After this paper had been published, we discovered a similar model: SSN-LDA [13]. Although the models are formally similar, their motivations and applications are different.

graphs (e.g., social networks, citation networks, etc) and text corpora. The first difference is the distribution of node degrees vs. distribution of words. Text corpora usually have at least tens or hundreds of words in even the shortest documents. Social networks tend to have power-law distributions [9], with the modal degree being 1 and the mean degree being very low. The second major difference is that many real-world graphs, especially social networks, are not represented as multigraphs. Clearly, most documents of interest have some words that appear multiple times; and LDA takes advantage of this when modeling the topics. Finally, text documents are semantically sensitive to word ordering, whereas in social networks, no order is assigned to the edges of a given vertex.

Fortunately, only one of these differences needs to be considered in developing LDA-G. Word order is not part of the LDA model, so it can be disregarded in our adaptation. The multigraph aspect is a valid consideration; but in our experiments (see Section 5), it does not have any major effect on performance. In fact, when we analyzed a real-world multigraph using LDA-G, it generated a more predictive model when all multiple edges were transformed into single edges. Thus, the multigraph discrepancy can be safely ignored. The degree distribution, however, does have an effect on performance. In particular, when there are hundreds of thousands of low-degree nodes in a network, the Gibbs sampler for LDA tends to generate thousands or tens of thousands of groups. The reason for this can be readily seen by looking at the sampling equations in [4]. This is a problem for group discovery, since these extra "topics" or groups have no significant meaning for real-world graphs, and do not provide additional predictive capacity on the topology of the graph.

This obstacle leads to the primary difference between LDA and LDA-G. To perform group discovery on graphs, LDA-G can use either a *finite* parametric model or an *infinite* nonparametric model. In the finite case, the model is truncated at some number of groups, K, which is chosen on a per-application basis. There are several considerations which can affect the proper choice of K. For our purposes, we use link prediction on a tuning set as the main criterion. The infinite LDA-G model learns the appropriate model size through yet another hyperparameter and alterations to the Gibbs sampler update. In our experiments (see Section 5), we find finite LDA-G with model selection to be superior to the full nonparametric version.

Given our choice of the finite LDA-G model, we also make another modification, which is motivated by our own empirical observations rather than any characteristics of real-world graph data. This alteration pertains to a more-careful choice for the starting configuration of the Gibbs sampler. We present results in Section 5 that demonstrate the drastic effect the initial configuration has on link prediction results. Previous work on Gibbs sampling for LDA [4] suggest that any starting configuration is acceptable. We find that the best starting configuration is to assign topics in a random order, but to condition each initial assignment on all previous assignments using the Gibbs sampling mechanism.

As stated in the previous section, LDA-G's runtime complexity is $O(NKM)$; its space complexity is $O(N(K+M))$. Here, N is the number of vertices in the input graph, K is the number of groups (where $K \ll N$), and M is the average vertex degree in the graph. In real-world graphs, M scales at most logarithmically (and not as a power law) with N [7]. These complexity requirements are much better than IRM's $O(N^2)$ complexity for runtime and space. Moreover, LDA-G can be readily applied to large graphs with hundreds of thousands of nodes by utilizing David Newman *et al.*'s work [8] on distributed inference for LDA.

4. EXPERIMENTAL STUDY

For our experimental study, we compare LDA-G with IRM and Cross-associations. We use three datasets collected from PubMed[4] in our analysis (see Table 4). They are as follows.

Author x Knowledge Graph: This is a bipartite graph with author nodes connected to knowledge theme[5] nodes. A link exists from an author u to a knowledge theme k for every paper in which u is a coauthor and k is a theme appearing in the abstract. This graph is a multigraph, but as noted above we summarize any multiple edges as single edges. This graph contains 37,346 authors and 117 knowledge themes, with 119,443 author-knowledge edges (see Figure 2).

Author x Author Graph: This graph is a coauthorship network, two author nodes are connected if they published a paper together. This symmetric graph has a clique for each publication. It contains 37,227 authors and 143,364 edges (see Figure 2). Note that there are some authors who appear in this graph but not in the Author x Knowledge Graph and vice versa.

Knowledge-Infused Author x Author Graph: This graph is constructed by infusing the Author x Author Graph with information from the Author x Knowledge Graph. We generate this graph as follows. First, we prune the Author x Knowledge multigraph by removing links that appear less than 12 times. We pick the threshold of 12 based on graph-size considerations. This step effectively removes noise in the Author x Knowledge Graph. Second, in the Author x Author Graph, we add a link between any pair of authors that share a knowledge theme in the pruned Author x Knowledge graph. This produces a denser "knowledge-infused" Author x Author Graph with 37,227 authors and 339,644 edges (see Figure 2).

Sampled Graphs: Because IRM cannot process graphs larger than a couple of hundred nodes, we also create sampled versions of each network. For the Sampled Author x Knowledge Graph, we simply select 50,000 entries at random from the adjacency matrix. For the Sampled Author x Author Graph, we select 100 authors at random and retained their entire edge-lists (including authors not present in the selected 100). For the Sampled Knowledge-Infused Author x Author Graph, we add an edge to the Sampled Author x Author Graph whenever two authors from the selected 100 share a neighbor in the Author x Knowledge Graph.

Given the aforementioned sets of graph data and algorithms, our task is to find meaningful latent groups, and then use these groups to find teams that are performing research similar to [6] and [11].

[4] PubMed is a repository containing millions of citations from biomedical articles (see http://www.pubmedcentral.nih.gov/).

[5] Knowledge themes were extracted based on term frequency in PubMed abstracts.

Table 1: Graph datasets used in our evaluations of LDA-G and Cross-associations. IRM cannot process graphs this large, so it is evaluated on sampled versions of each graph.

Input Graph	Number of Nodes	Number of Links
Author x Knowledge	37,346 authors & 117 knowledge-themes	119,443
Author x Author	37,227 authors	143,364
Knowledge-Infused Author x Author	37,227 authors	339,644

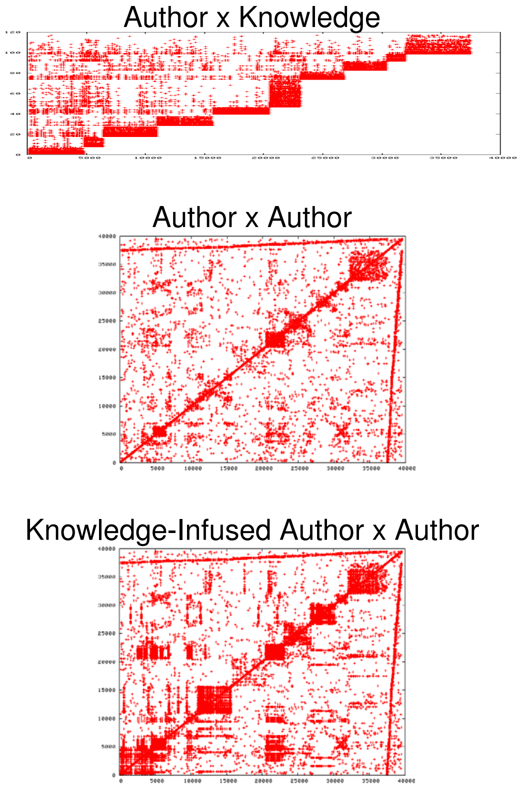


Figure 2: Adjacency matrices for our data graphs. (Note: Author x Knowledge Graph's adjacency matrix is sorted by the order in which each author's knowledge theme was added to the graph.)

5. RESULTS

In this section, we discuss the quantitative and qualitative results of our experiments.

5.1 Quantitative Results

We measure the superiority of a group discovery algorithm by how well its groups predict the topology (i.e. link structure) of the graph. To this end, we use link prediction as a metric.

As is standard in machine learning, we divide the dataset into training and test sets, build a model on the training set, and examine its performance with respect to the chosen metric on the test set. In particular, we use stratified random sampling to hold-out some links from the input graph. The remaining links are used to discover the latent groups. Then, the superiority of these groups is checked based on how well they predict the existence of the held-out links.

Each algorithm is able to assign a probability $p(v_i \rightarrow v_j)$ to the existence of an edge between vertex v_i and vertex v_j. For IRM, this is the average over all sampled configurations of the element η_{v_i,v_j}. For Cross-associations, it is the link density of the block that the edge falls under. For LDA-G, it is the average over all sampled configurations of $\sum_{t \in topics} p(t|v_i)p(v_j|t)$.

In all experiments, we construct the ROC curve on a test set of *unobserved* links (randomly selected from the graph data) and report the average Area Under the ROC Curve (AUC) over 5 independent trials.

For the finite LDA-G model, we bound the number of groups at 34. Our model selection is based on link prediction results from a separate tuning set for each dataset.

5.1.1 *Author x Knowledge Graph*

To test LDA-G and Cross-associations, we use stratified random sampling to select 1000 links from the Author x Knowledge Graph. These 1000 links compose our test set. The remaining 118,443 links in the Author x Knowledge Graph form our training set and are used to discover the latent groups in the graph.

IRM cannot handle a graph as large as the Author x Knowledge Graph (with approximately 37K nodes and 119K links), so we use stratified random sampling to select 1000 links[6] from the Sampled Author x Knowledge Graph (described in Section 4). The remaining 49,000 links (from the Sampled Author x Knowledge Graph) are used (by IRM) to find the hidden groups in the graph.

Figure 3 summarizes the link prediction results on the Author x Knowledge Graph. Even with 99% of the data discarded, IRM is very predictive, averaging 0.863 AUC over 5 trials. Cross-associations averages 0.914. LDA-G dominates both methods with an average AUC of 0.955. We also used LDA (unmodified) to analyze this graph; its average AUC was 0.777 on the same test sets. This demonstrates the drastic effect of the minor modifications that transform LDA into LDA-G.

In addition, we tested LDA-G with three different starting configurations on this data. The results are summarized in Figure 4. The *Random* configuration simply assigns each edge to a random topic initially. The *CRP* configuration uses a Chinese Restaurant Process [1], and the *Conditioned* configuration is as described in Section 3. As expected, a good initial configuration improves LDA-G's performance.

5.1.2 *Author x Author Graph*

To test LDA-G and Cross-associations on the Author x

[6] We select both present (1-valued) links and (absent 0-valued) links.

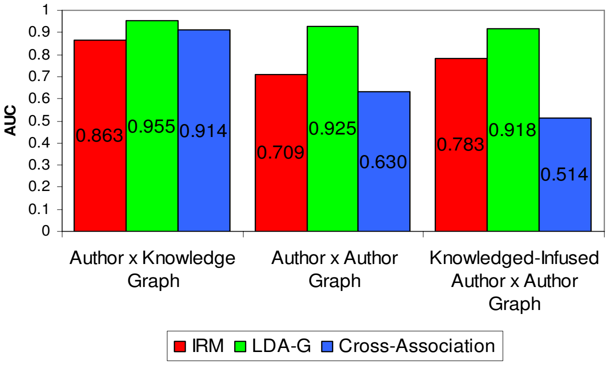


Figure 3: Area under the ROC curve for link prediction (averaged over 5 trials). An AUC of 0.5 is a random guess.

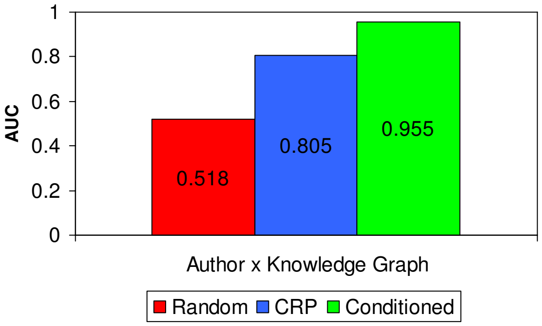


Figure 4: Area under the ROC curve results on link prediction from three different starting configurations for LDA-G (averaged over 5 trials)

Author Graph, we randomly select 500 present (1-valued) edges and 500 absent (0-valued) edges to test link-prediction on. The remaining 142,364 edges are used to find groups.

Once again, IRM cannot handle a graph as large as the Author x Author Graph (with 37K nodes and 143K links), so we randomly select 10% of the edges (present and absent) from the Sampled Author x Author Graph (see Section 4) and train on the remaining 90% edges from the Sampled Author x Author Graph.

Figure 3 depicts the results. On the entire Author x Author Graph, we see the superiority of LDA-G even more clearly. IRM fails to discover any meaningful groups, and Cross-associations appears to "give up" after a few iterations without having done much reordering at all.

5.1.3 *Knowledge-Infused Author x Author Graph*

The experimental methodology for the Knowledge-Infused Author x Author Graph is identical to the one used with the Author x Author Graph in the previous section. The results are listed in Figure 3.

For IRM, these results demonstrate the value of fusing different types of data. The average AUC jumps from 0.709 in the Author x Author Graph to 0.783, indicating that the structured Author x Knowledge data can "fill in the gaps" in the coauthorship data. The effect on LDA-G is a small decrease in AUC, most likely due to the fact that LDA-G predicts the topology of the Author x Author Graph very reliably, and the extra information serves to confuse the model slightly. The failure of Cross-associations here is total; it performs only marginally better than a random link prediction model.

5.1.4 *A Brief Note on Running Time*

While a full comparison of time and space complexity exceeds the scope of this paper, we will briefly summarize our experiences with the three algorithms under consideration. Note that IRM (due to its inability to scale to large graphs) is running on fundamentally different (and much smaller) datasets than the other two algorithms, but we will report its performance nonetheless.

In experiments with the Author x Knowledge Graph, LDA-G and IRM take around 15 minutes to perform inference on a standard commodity desktop with 2Gb memory. We assume that 50 scans of the data is sufficient in all Gibbs sampler chains. Cross-associations completes in about 90 minutes here. For the Author x Author Graph, all three approaches complete in around 20 minutes. For the Knowledge-Infused Author x Author Graph, LDA-G and IRM each require about 45-60 minutes, while Cross-associations gives up after 10 minutes. Recall that in this experiment, Cross-associations performs slightly better than random group assignment.

5.2 Qualitative Results

We have established that LDA-G discovers groups which are highly predictive of the underlying topology of the graphs (as measured by link prediction), so we can expect to find useful qualitative results if we examine these groups. All of the qualitative results are based on the maximum likelihood configuration.

Figure 5 depicts LDA-G's qualitative results on the Author x Knowledge Graph. Even though Kobasa *et al.*'s paper [6] and Tumpey *et al.*'s paper [11] are both on the reconstructed 1918 influenza virus, the probability distribution for knowledge themes of major author groups of [6] and [11] are different.

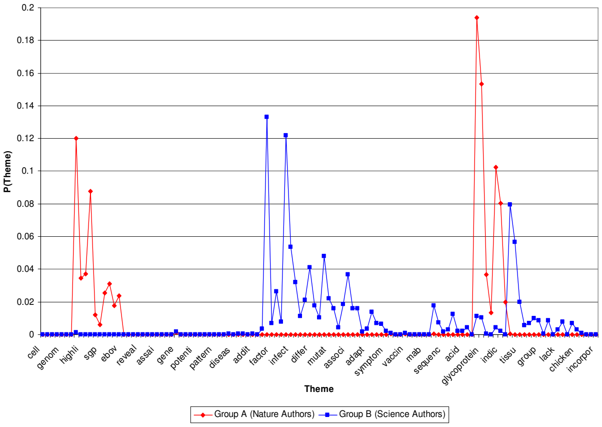


Figure 5: LDA-G's qualitative results on the Author x Knowledge Graph: Probability of knowledge-themes for major author groups of [6] in red and [11] in blue

In the Author x Author Graph, LDA-G finds one group that is common between the authors of [6] and [11]. In particular, 83% of the authors in [11] and 9% of authors in [6] fall in this group (see Figure 6, top plot, group 6). Further investigation of this group reveals other authors that have similar coauthorship patterns as the authors of [6] and [11].

In the Knowledge-Infused Author x Author Graph, LDA-G discovers three common groups between [6] and [11]. In particular, one of the groups includes 47% of authors in [11] and 30% of authors in [6] (see Figure 6, bottom plot, group 3). Looking into this group reveals other authors that have both similar coauthorship patterns and knowledge themes as authors of [6] and [11]. As expected, fusing data from different sources, especially introducing more structured data into less structured graphs, can be quite valuable.

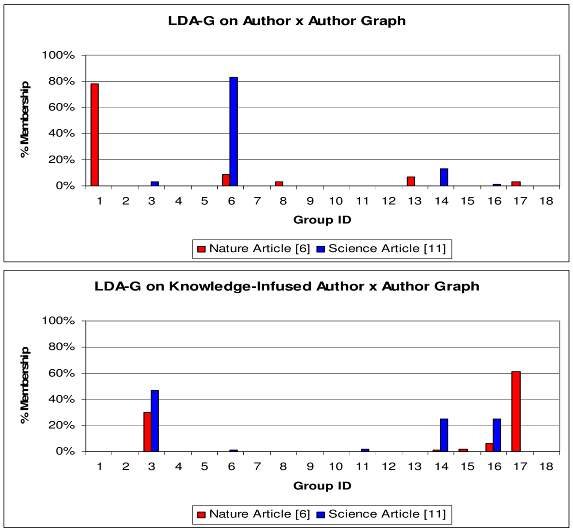


Figure 6: LDA-G's qualitative results on the Author x Author Graph and the Knowledge-Infused Author x Author Graph. The latter graph (bottom plot) generates groups with larger overlap among the authors of [6] and [11] than the former graph (top plot). (Note: The group IDs between the two plots do not refer to the same groups.)

6. CONCLUSIONS AND FUTURE WORK

Group discovery in graphs is a challenging problem. Most existing solutions are either computationally scalable but not expressive (in terms of their group representation), or are expressive but not scalable. This paper describes LDA-G, a scalable and expressive approach to this problem. LDA-G modifies a popular topic modeling algorithm for graph data. Comparative experiments with IRM (a Bayesian approach) and Cross-associations (a compression-based approach) illustrate the superiority of LDA-G in terms of both quantitative and qualitative results. In particular, IRM fails to scale to graphs with thousands of nodes, while Cross-associations fails to produce meaningful groups on two out of the three graph datasets (namely, the Author x Author Graph and knowledge-infused Author x Author Graph).

In our experimental study, we also demonstrate that fusing data from different sources, especially introducing more structured data into less structured graphs, can improve quantitative and qualitative results significantly. This has implications not only for knowledge discovery algorithms, but also for data collection efforts.

Future work includes parallelizing inference in LDA-G (similar to [8]). Also, we would like to modify LDA-G to handle multiple types of vertices and edges, as well as attributes on vertices and edges. Many real-world graph data sets have these properties, and we believe they will improve both quantitative and qualitative performance of LDA-G. Lastly, we are exploring a temporal version of LDA-G, which tracks the evolution of discovered groups over time in dynamic graph data.

7. REFERENCES

[1] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. *NIPS*, 16, December 2003.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[3] D. Chakrabarti, S. Papadimitriou, D. Modha, and C. Faloutsos. Fully automatic cross-associations. In *Proc. of the 10th KDD*, pages 79–88, 2004.

[4] T. Griffiths. Gibbs sampling in the generative model of latent Dirichlet allocation. Technical report, Stanford University, 2002. Available at http://citeseer.ist.psu.edu/613963.html.

[5] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *Proc. of the 21st AAAI*, pages 381–388, 2006.

[6] D. Kobasa, S. Jones, K. Shinya, J. Kash, J. Copps, H. Ebihara, Y. Hatta, J. Kim, P. Halfmann, F. Feldmann, J. Alimonti, L. Fernando, M. Katze, H. Feldmann, and Y. Kawaoka. Aberrant innate immune response in lethal infection of macaques with the 1918 influenza virus. *Nature*, 445:319–323, 2007.

[7] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM TKDD*, 1(1), 2007.

[8] D. Newman, A. Asuncion, P. Smyth, and M. Welling. Distributed inference for latent Dirichlet allocation. *NIPS*, 20:1081–1088, December 2007.

[9] M. E. J. Newman. The structure and function of complex networks. *SIAM Rev.*, 45(2):167– 256, 2003.

[10] M. E. J. Newman. Modularity and community structure in networks. *PNAS USA*, 103:8577, 2006.

[11] T. M. Tumpey, C. F. Basler, P. V. Aguilar, H. Zeng, A. Solorzano, D. E. Swayne, N. J. Cox, J. M. Katz, J. K. Taubenberger, P. Palese, and A. Garcia-Sastre. Characterization of the reconstructed 1918 spanish influenza pandemic virus. *Science*, 310(5745):77–80, 2005.

[12] Z. Xu, V. Tresp, S. Yu, K. Yu, and H.-P. Kriegel. Fast inference in infinite hidden relational models. In *Proc. of the 5th Int'l Workshop on Mining and Learning with Graphs*, 2007.

[13] H. Zhang, B. Qiu, C. L. Giles, H. C. Foley, and J. Yen. An lda-based community structure discovery approach for large-scale social networks. In *Proc. of the IEEE Int'l Conf. on Intell. and Security Informatics*, pages 200–207, New Brunswick, NJ, May 2007.